IBM Data Science Professional Certificate – Applied Capstone Project

Depicting Country-wise population distribution over recommended venues near London

Roudrav Chakraborty

28th July, 2020

Table of Contents :

1. Introduction

1.1 Background

London, the capital of England and the biggest city in the entire United Kingdom is probably one of the most famous cities in the world. Also known as “The Swinging City” and “Home of The Big Ben”, it surely is a collection of magnificent places. From ‘Buckingham Palace’ and ‘The Big Ben’ to ‘Westminster Abbey’ and ‘The London Eye’, the city is a galore of attraction points.

That being said, there are always recommended spots in a city that attract people of all tastes and get tagged as ‘Must Go-to’ areas. People from all corner of the city tend to check them out and get to explore the various aspects of these venues. The city being host to a diverse collection of people from different nationalities, enables people from all around the globe to experience the areas.

It would be interesting to see how country-wise trend (outside United Kingdom) in people going to a particular recommended place has changed over a particular span of time. This data if made available to those places in a visual manner, might allow the respective venues to improve on their qualities by seeing the trend for different countries. This can lead to higher revenues and might allow the businesses to flourish.

1.2 Problem

The count of the people in the city segregated country-wise and coupled with the likes for a recommended place can give a fairly good idea (by assumption of a scale) about the distribution of the population over those venues. Hence this project aims to depict the country-wise distribution of the population over a recommended area and

visualize the change (in trend of people from different nationalities visiting) over a period of time, which when used would ensure that opportunities for improvement or new ideas come up for those places.

1.3 Interest

The visualized data, when made public, would be useful to the recommended places as they can analyze the factors behind their success. It would also be helpful for currently under-performing areas to get delve deeper and find out as to what elements impact a nationality's preference of a place.

2. Data acquisition and cleaning

2.1 Data sources

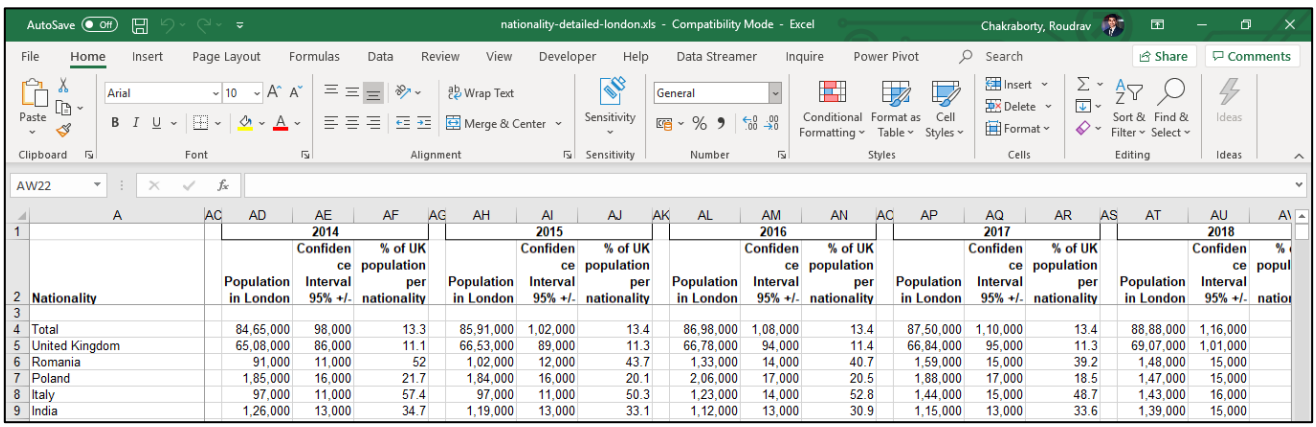
The data was obtained from the London Datastore which contained the **Detailed Nationality** dataset showing the nationality estimates for specific countries outside United Kingdom such as India, United States, Australia, Portugal, France etc. available for London as a whole. The data set contains the population residing in London per country across 12 years. However, we will be working with the following 5 nationalities for a period of 2008-2018 i.e., 10 years :

- India
- France
- United States
- Portugal
- Australia

Apart from this, we would be leveraging the Foursquare API endpoints to obtain the **Recommended** places in London at the time of execution. This data would be coupled with the **Likes** per venue to give a brief idea as to how many people would be visiting a recommended place. Upon cross-referencing them with the population dataset, we would depict a pattern of nationalities towards those places.

2.2 Data cleaning

The population data was extracted into pandas dataframe **nat_df**. Firstly, the dataset had multiple columns merged into a single cell, hence they were not labelled properly. The year columns were the only headers that were correctly marked. Below is an example of how the dataset looked like :



	A	AD	AE	AF	AH	AI	AJ	AL	AM	AN	AP	AQ	AR	AT	AU
1		2014			2015			2016			2017			2018	
2	Nationality	Population in London	Confidence Interval 95% +/-	% of UK population per nationality	Population in London	Confidence Interval 95% +/-	% of UK population per nationality	Population in London	Confidence Interval 95% +/-	% of UK population per nationality	Population in London	Confidence Interval 95% +/-	% of UK population per nationality	Population in London	Confidence Interval 95% +/-
3															
4	Total	84,65,000	98,000	13.3	85,91,000	1,02,000	13.4	86,98,000	1,08,000	13.4	87,50,000	1,10,000	13.4	88,88,000	1,16,000
5	United Kingdom	65,08,000	86,000	11.1	66,53,000	89,000	11.3	66,78,000	94,000	11.4	66,84,000	95,000	11.3	69,07,000	1,01,000
6	Romania	91,000	11,000	52	1,02,000	12,000	43.7	1,33,000	14,000	40.7	1,59,000	15,000	39.2	1,48,000	15,000
7	Poland	1,85,000	16,000	21.7	1,84,000	16,000	20.1	2,06,000	17,000	20.5	1,88,000	17,000	18.5	1,47,000	15,000
8	Italy	97,000	11,000	57.4	97,000	11,000	50.3	1,23,000	14,000	52.8	1,44,000	15,000	48.7	1,43,000	16,000
9	India	1,26,000	13,000	34.7	1,19,000	13,000	33.1	1,12,000	13,000	30.9	1,15,000	13,000	33.6	1,39,000	15,000

Hence, I first generated a list containing the first column (which was the country name) and the 10 consecutive years vis-à-vis 2008 to 2018. From the entire dataset, I extracted the data for only these 11 columns.

Next, due to blank cells at the top of the sheet and the bottom of the sheet, a lot of NaN values were present. So the first 2 records and the last 3 records were dropped from the dataframe to ensure that only the respective county names are present along with the years. Then, I set the 'Country' column as the index of the dataframe so that it becomes easier to slice and data retrieval is simplified.

Here is how the cleaned dataframe looked like :

nat_df

Country	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
India	99000	127000	128000	156000	131000	126000	119000	112000	115000	139000
France	68000	57000	81000	68000	79000	91000	88000	100000	92000	84000
United States	43000	42000	46000	53000	49000	50000	56000	48000	51000	56000
Portugal	54000	42000	53000	56000	64000	73000	86000	86000	93000	76000
Australia	41000	35000	38000	37000	38000	48000	37000	43000	40000	39000
Total Population of London	7870000	7985000	8130000	8236000	8341000	8465000	8591000	8698000	8750000	8888000

Next, I needed the recommended venues within a specific range of the city of London. I defined a custom radius of 50000 meters (50 Kilometers) for extracting the data and fetched up to a limit of 100 venues. The response of the **Foursquare Explore** endpoint was JSON collection which required to be wrangled in order to extract meaningful information. It was the same case for **Foursquare Likes** endpoint, which returned the data in a JSON format.

In both cases, the lists and dictionaries in the JSONs were indexed and the required information was extracted. The dataframe **venues_df** for Venues was created first and then a list containing the Like counts was appended to the dataframe as a column. The dataframe was then sorted in descending order on the basis of the Like count and the **top 13 rows (corresponding to the top 13 recommended places)** were stored.

The final cleaned dataframe looks as follows :

venues_df

	Venue_ID	Venue_Name	Latitude	Longitude	Category	Likes
0	4ac518d2f964a52026a720e3	Hyde Park	51.507781	-0.162392	Park	11636
1	4ae5b238f964a52087a121e3	Selfridges & Co	51.514640	-0.152864	Department Store	9163
2	4ac518d2f964a5203da720e3	British Museum	51.519009	-0.126437	History Museum	8539
3	4ac518eff964a52064ad20e3	Borough Market	51.505495	-0.090518	Farmers Market	7590
4	4ac518cef964a520f6a520e3	Big Ben	51.500620	-0.124578	Monument / Landmark	6467
5	4ac518cef964a520f7a520e3	Tower of London	51.508248	-0.076261	Castle	4876
6	4ac518cdf964a520e6a520e3	National Gallery	51.508876	-0.128478	Art Museum	4689
7	4ac518cef964a520f9a520e3	Trafalgar Square	51.507987	-0.128048	Plaza	4531
8	4ba6419bf964a520b23f39e3	Covent Garden Market	51.511977	-0.122799	Shopping Plaza	3556
9	4b233922f964a520785424e3	Regent's Park	51.530479	-0.153766	Park	3446
10	4bd2fa6641b9ef3becf7fee5	Old Spitalfields Market	51.519668	-0.075375	Flea Market	3233
11	4ac518cdf964a520f2a520e3	St James's Park	51.503253	-0.132995	Park	3131
12	4ac518ecf964a52072ac20e3	Liberty of London	51.513725	-0.140510	Department Store	2695

2.3 Feature selection

For the Nationality dataframe, the data was selected for the following countries :

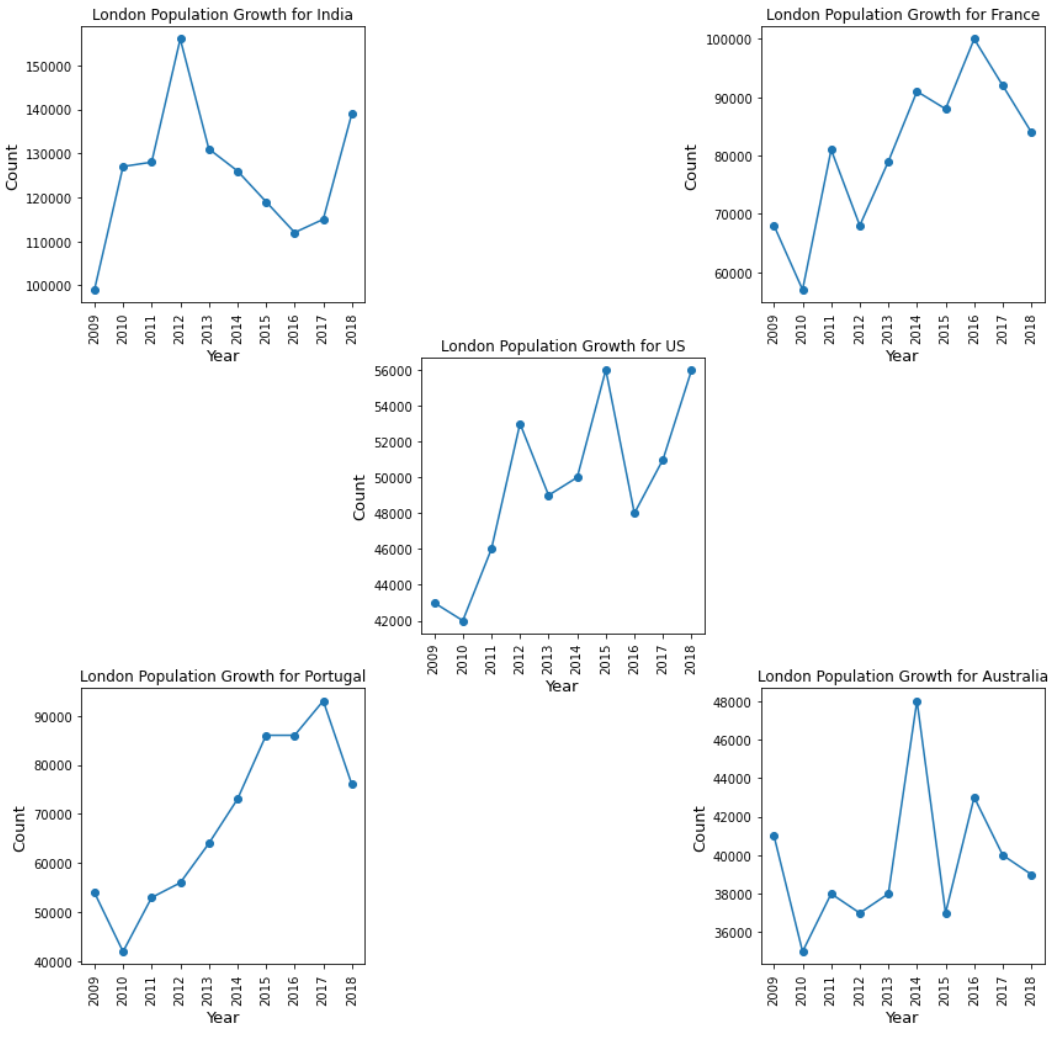
- India
- France
- United States
- Portugal
- Australia

The year range was selected as 2008 – 2018 to allow proper non-cluttered visualization. For the Venues dataframe, unnecessary columns such as address, referrals etc. were excluded and only the relevant fields such as Venue ID, Venue Name, Coordinates (Latitude and Longitude) and Category were kept. A list of Likes per venue was added as a column in the Venue dataframe.

3. Methodology

We start by depicting the growth of the population per country over the selected period of years. This will give us an idea of the chances of people from a given nationality to a particular recommended place. We transpose the original nationality dataframe **nat_df** to obtain the year values as a column. Then we create a subplot grid of size 3x3 and plot the data of 5 countries in that.

The figure is observed as follows :



Now, we plot the venues over the map of London to get a brief idea of the vicinity of each place and whether they can impact each other's business. The map is plotted using the Folium library and circle markers with popup are placed to show the location of each venue as follows :

Next, we set the Year back as index and then calculate the percentage of different nationalities with respect to the population of London. This population when multiplied with the scaled Like count can be a metric to estimate the trend of different nationalities at a given place. The percentage dataframe **nat_df_tran_per** looks like this :

```
# Displaying the percentage dataframe
nat_df_tran_per
```

	Year	India	France	US	Portugal	Australia	Total Population of London
0	2009	0.0125794	0.00864041	0.00546379	0.0068615	0.00520966	1
1	2010	0.0159048	0.00713838	0.00525986	0.00525986	0.00438322	1
2	2011	0.0157442	0.0099631	0.00565806	0.00651907	0.00467405	1
3	2012	0.0189412	0.00825644	0.00643516	0.00679942	0.00449247	1
4	2013	0.0157056	0.00947129	0.0058746	0.00767294	0.00455581	1
5	2014	0.0148848	0.0107501	0.00590667	0.00862374	0.00567041	1
6	2015	0.0138517	0.0102433	0.00651845	0.0100105	0.00430683	1
7	2016	0.0128765	0.0114969	0.00551851	0.00988733	0.00494367	1
8	2017	0.0131429	0.0105143	0.00582857	0.0106286	0.00457143	1
9	2018	0.0156391	0.00945095	0.00630063	0.00855086	0.00438794	1

Now, we make an important assumption. We scale the venue likes according to the average population of the city over the last 10 years. We have **assumed** that the number of likes given per venue has been from a sample of 50000 users. So, if a venue has 5000 likes, then according to our assumption among 50000 users who visited the place, 5000 explicitly posted a like for it, which was eventually captured in the Foursquare database.

Based on this assumption, we scale the Likes column of the **venues_df** dataframe by multiplying the Like count per place with the mean population of London over the last 10 years and then dividing it by out sample size i.e., 50000. It is then rounded off to the nearest lower whole number.

The new **venues_df** dataframe with the updated like count looks as follows :

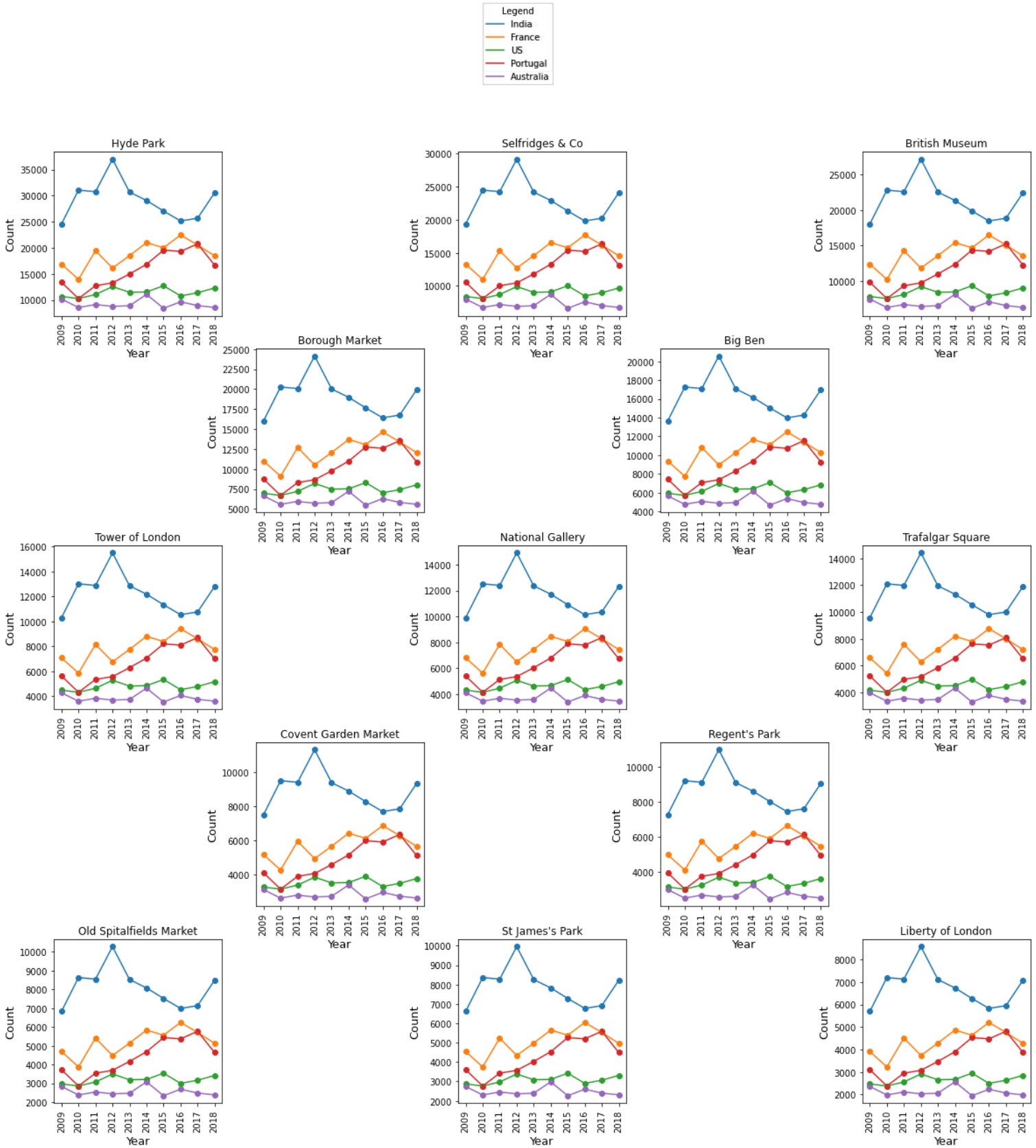
venues_df

	Venue_ID	Venue_Name	Latitude	Longitude	Category	Likes
0	4ac518d2f964a52026a720e3	Hyde Park	51.507781	-0.162392	Park	1953777.0
1	4ae5b238f964a52087a121e3	Selfridges & Co	51.514640	-0.152864	Department Store	1538541.0
2	4ac518d2f964a5203da720e3	British Museum	51.519009	-0.126437	History Museum	1433766.0
3	4ac518eff964a52064ad20e3	Borough Market	51.505495	-0.090518	Farmers Market	1274421.0
4	4ac518cef964a520f6a520e3	Big Ben	51.500620	-0.124578	Monument / Landmark	1085861.0
5	4ac518cef964a520f7a520e3	Tower of London	51.508248	-0.076261	Castle	818719.0
6	4ac518cdf964a520e6a520e3	National Gallery	51.508876	-0.128478	Art Museum	787320.0
7	4ac518cef964a520f9a520e3	Trafalgar Square	51.507987	-0.128048	Plaza	760791.0
8	4ba6419bf964a520b23f39e3	Covent Garden Market	51.511977	-0.122799	Shopping Plaza	597080.0
9	4b233922f964a520785424e3	Regent's Park	51.530479	-0.153766	Park	578610.0
10	4bd2fa6641b9ef3becf7fee5	Old Spitalfields Market	51.519668	-0.075375	Flea Market	542846.0
11	4ac518cdf964a520f2a520e3	St James's Park	51.503253	-0.132995	Park	525719.0
12	4ac518ecf964a52072ac20e3	Liberty of London	51.513725	-0.140510	Department Store	452512.0

Next, we plot a collection of 13 scatter plots (corresponding to 13 different places) on a 5x5 subplot grid to depict the trend of different nationalities in one single go. Each scatter plot contains the data for all the 5 selected countries over a period of 10 years. The Matplotlib library has been used to generate the subplots and the Pyplot layer along with Axes method has been used to generate the scatter plots.

To prevent excessive lengthy code, we have generated an Axes list which store the subplot information. This Axes list is referred inside two for loops where the Like count for each individual venue is multiplied by the population percentage (as obtained above) to get the rough count of a nationality's presence at a particular place.

The final scatter plot collection looks as follows :



4. Results

After conducting the study and checking variety of details, the following are the results :

- People tend to focus more on historical areas, department stores, shopping areas and markets rather than restaurants, performing art venues and bookstores.

- Indians and Americans have shown considerable interest in checking out the recommended areas, which can come from the fact that their population has seen a significant rise over the past few years.

- Since restaurants, concert halls and coffee shops are not at the top of the recommendation list, these businesses need to revisit their plans to come up with more attractive offers for better sustainability.

- World famous places such as Big Ben, National Gallery, British Museum etc. have not been the most recommended places. Therefore, we can safely assume that all places need to upgrade themselves with time to stay within recommendations.

5. Discussion

Based on the data and the charts plotted accordingly, we can infer the following :

- Hyde park is one of the most recommended areas in the city. A further internet search portrays that it is the largest of four Royal Parks, which explains its popularity among people from all countries.

- Department Stores and Parks are the major categories under which most of the recommended places fall.

- Most of the recommended places are centered around the river Thames, which means that for a potential new businessperson, the area around River Thames could be a major spot.

- In general, most of the recommended places are situated at a relatively similar distance from each other, which means that business interference between two recommended places is minimal. This also comes from the fact that there are variety of categories present for the recommended places.

- 3 out of 5 countries under study have shown a decline in the population in London. This can be due to many reasons, but it cannot be ruled out that the city's recommended places are not attracting them anymore.

6. Conclusion

As a part of this project and study, we demonstrated that with the right dataset and the right skills, plausible relationships can be established between features which can lead to deeper insights. We studied the trend of population changes for people of 5 nationalities over a period of 10 years for recommended places in London area.

These studies can be vital for businesses to identify loopholes or diminishing traits due to which they are not able to hold on to their customer/tourist base of various cultural and national diversities. On the other hand, it also allows the flourishing businesses to analyze their success factors and improve them further for a better experience.

This concludes the report for this Capstone project. Thank you.